DOCUMENT RESUME

ED 306 311 TM 013 265

AUTHOR Brauchle, Paul E.; And Others
TITLE Developing and Evaluating a Method of Using Student Performance and Attitude Data as a Measure of Teacher Effectiveness.
PUB DATE Apr 87
NOTE 20p.; Paper presented at the Annual Meeting of the National Council for Measurement in Education (Washington, DC, April 21-23, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Attitude Measures; Career Ladders; Educational Assessment; Elementary Secondary Education; *Evaluation Methods; Interrater Reliability; Public School Teachers; *Rating Scales; State Programs; *Student Attitudes; *Teacher Effectiveness; Teacher Evaluation; Test Construction; Testing Programs
IDENTIFIERS Tennessee Career Ladder Program

ABSTRACT

The results of the first year of study of an approach in Tennessee using student performance assessments in the evaluation of teachers are summarized and the results of a field test on student performance and attitude are presented. The Tennessee approach places the responsibility for selecting and presenting the data on the teachers. About 130 trained evaluators in individual face-to-face meetings with teachers attempted to ascertain the teacher's performance on two items they presented: (1) evidence of student academic achievement commensurate with student background/ability; and (2) evidence of improvement in student attitude toward learning. Results from a sample of 201 teachers--out of the 1,745 Tennessee Career Ladder teachers who were evaluated in 1985-86--were used. There was limited reliability in the measurement of the two items since evaluators' ratings on them differed widely for the same teachers. The evaluators appeared to have developed unique operational definitions of the scale points. The scales were valid in that the scale point definitions suggested by evaluator comments on the dialogues matched the semantically developed anchor point definition. Based on these results, an operational definition for five scale anchor points was constructed. Findings suggest that the performance items should not be used in their present form as part of the Career Ladder scores for teachers, but should be continued as field test items. Evaluators should receive specific training in the meaning and definitions of the scale points and how to reliably assess documentation on student performance. If evaluators are adequately trained, student performance and attitude data as presented by teachers can become a reliable part of a teacher evaluation system. (SLD)

ED306311

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. FERGUSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Developing and Evaluating a Method of Using Student Performance and Attitude Data as a Measure of Teacher Effectiveness

Paul E. Brauchle
Tennessee Department of Education

Joyce McLarty
American College Testing Program

James Parker
Tennessee Department of Education

Paper Presented at the Annual Meeting
of the National Council for Measurement
in Education, Washington, D.C., April, 1987

Running Head: TEACHER EFFECTIVENESS

TM013265

Teacher Effectiveness

The use of student performance as an index of teacher effectiveness is not new. It has been seriously studied for at least four decades (Bolin, 1984) and has engendered at least a moderate degree of controversy (Ornstein, 1983; Reiff, 1981; NJEA, 1978; NJEA, 1979; Medley, 1985; and Soar and Soar, 1975). It has been used in several curriculum areas (Cochran and Mills, 1983; Patterson and Keel, 1979), and at various levels (RELCV, 1970; Johnson, 1982; Capie, et.al., 1980; Centra, 1979) with various degrees of success (Morrison, 1974; Schein, 1985). It has been proposed as a basic measure of school accountability (Genck, 1982), used as an index of student teacher growth (Denton and Norris, 1979), and of teacher performance (Lucio, 1973; Hall, 1981); and has been suggested as a basis of a teacher merit pay system (Berk, 1984). It has, of course, been used as a criteria against which to validate other measures of teacher performance (James, 1978; Capie, 1980; Ellett, et.al., 1980).

In 1978, Bartz, et.al., proposed student achievement as one component in a multiple approach to teacher evaluation. Today, Tennessee's Career Ladder system for the merit evaluation of teachers uses student performance and student attitudes toward learning as components of the teacher evaluation system, although at present they are field test items and are not scored.

During 1985-86, the State of Tennessee pilot tested a means of utilizing student performance data in the evaluation of teachers for the Career Ladder program. Unlike other approaches to using student data in assessing teachers, the Tennessee approach places the responsibility for selecting and presenting the data on the teacher. Tennessee teachers undergoing the Career Ladder evaluation process were required to submit evidence of improvements in student academic achievement and student attitude toward learning. This evidence was evaluated by State selected and trained evaluators. In addition, students were asked about improvements in their own achievements and attitudes, and principals were asked to judge the extent to which each teacher improved the academic achievements and attitudes of his or her students.

This paper summarizes the results of the first year of study of this approach to using student performance in the evaluation of teachers, and presents the results of the qualitative examination of field test data on student performance and attitude. It examines the merits and challenges of this approach to the utilization of student performance data in teacher evaluation. It also describes changes in evaluator training and rating scale construction used during the 1986-87 year. Some of the changes involve modifications of the training given to evaluators who will collect data from teachers and rate the teachers on the academic performance and attitude of their students.

Statement of Problem

The 1985-86 Tennessee Career Ladder contained student performance items on three different instruments, the student questionnaire, the principal questionnaire, and the teacher/evaluator dialogue--evaluation focus. Different student questionnaires were administered for each of the following levels: (1) grades K to 2, (2) grades 3 to 6, and (3) grades 7 to 12. The questionnaire for grades K to 2 did not contain any student performance items. The questionnaires for grades 3 to 6 and grades 7 to 12 contained two items each on student performance. The principal's questionnaire and the dialogue each contained two items on student performance. Of the two items on each of the instruments, one was related to student achievement and the other was related to student attitude toward education. These items were "field test" items during 1985-86, and were not used in the evaluation of teachers.

The two items were:

Item 11. Evidence of student academic achievement commensurate with background/ability of student; and,

Item 12. Evidence of improvement in student attitude toward learning.

The items were rated on a five point scale where 1=Unsatisfactory; 2=Below Average; 3=Average; 4=Outstanding; and 5=Distinguished.

The items were used in "dialogues", when trained evaluators in individual face-to-face meetings with teachers attempted to ascertain the teacher's performance on the items. In addition, the evaluators made written comments in support of their judgements of teacher performance on the items.

The problem was to determine what value the field test items have for evaluating teachers in the Tennessee Career Ladder system, and to specify if and how they could be equitably and fairly used as a part of a teacher's performance score.

Examination of the field test item data focused on six issues: (1) the content of the training provided for the approximately 130 state evaluators who collected the data, (2) the reliability of the data, (3) the adequacy of documentation provided by the candidates, (4) the nature of the documentation provided, (5) the results indicated by that documentation and (6) feedback provided by the evaluators and the candidates.

Method:

We obtained from a population of 1745 Career Ladder teachers who were evaluated in 1985-86 a random sample of sufficient size to represent that population at an error

rate of ± 5 percent, assuming a 80/20 split; that is, 80 percent of teachers had attained Career Ladder status. The required sample size was 201 teachers (Cochran, 1977). For each individual in the sample, we consulted the original dialogue record. We read the evaluator's notes, observed the evaluator's rating of the teacher on the five point scale, and recorded three things: (1) For each question, whether or not the comments of the evaluator would fit into a semantically developed conceptual scale of anchor points (see Footnote 1). (2) The evaluator's ratings on items #11 and 12, and (3) evaluation of the evaluator's comments and the ratings.

Results:

Based upon the examination of 201 randomly selected dialogue evaluations, the following results emerged:

(1) There was limited reliability across evaluators for written comments and ratings.

(2) There appeared to be systematic bias (high or low) for some evaluators.

(3) Observation data were often incorrectly reported as dialogue data.

(4) In some cases the ratings were not consistent with the written comments that the evaluators had made.

(5) Evaluators tended to look at process rather than performance. High ratings by the evaluators were often associated more with whether or not documentation existed than with whether or not a high level of student performance was indicated by the documentation.

(6) There were many anecdotal references that were difficult to tie to ratings because of their content.

Conclusions and Recommendations

We found that the two student performance items (Item #11 - Evidence of student academic achievement commensurate with background/ability of students; and Item #12 - Evidence of improvement in student attitude towards learning) did not appear to measure the same thing, since evaluators ratings on them differed widely for the same teachers. The written comments did not do a good job of supporting the scaled score. After comparing the written comments with the scaled scores, we concluded that the evaluators had developed their own unique operational definition of the scale points. However, the scales were considered to be valid in that the scale point definitions suggested by evaluator comments on the dialogues matched the semantically developed anchor point definition.

Based upon the relationship of comments to scaled scores for all evaluators over each item, the following operational definition for the scales anchor points seemed

appropriate: 1 = No information presented; 2 = Vague or unclear statements, no examples or pieces of evidence presented or evidence indicative with program; 3 = Clear statements with one or more specific examples or pieces of evidence presented. Evidence indicates average student progress; 4 = Clear statements with several examples or pieces of evidence presented. Evidence indicates above average student progress; and 5 = Clear statements with several examples or pieces of evidence presented, innovative procedures, extraordinary results or methods used, indicating exceptional student progress.

These findings suggested the following recommendations: 1) The two items should not be used in their present form as part of teachers' Tennessee Career Ladder scores. They should, however, be continued as field test items for the 1986-87 year.

2) Specific training should be given are to evaluators in terms of

- what the scale means,
- how each point is operationally defined.
- how to reliably assess documentation on student performance and student attitude.

3) The rating procedure for these items should be clearly documented in the training manual.

4) Examples of behaviors appropriate for each 1-5 rating should be given to evaluators and teachers.

Improving Evaluator Training

There are two specific goals of evaluator training for the student performance items. The first is to have different evaluators give the same rating to similar evidence presented by teachers. The second is to have all evaluators correctly rate student performance after having viewed the evidence.

The sequence of activities for training evaluators was similar to that used, for example, in working with a panel of experts except that evaluators were trained so that their evaluation consensus was not only reliable but was valid in that it was similar to that of experts for a given set of evaluation comments.

In training the evaluators for the 1986-87 year, actual comments from the 1985-86 dialogue which related to the student performance items were used. Comments were elicited from the evaluators for each anchor point and each instructional group. Those comments were used as reference points to operationally define each point on the scales for each item so that the evaluators knew conceptually and by example what kinds of evidence were appropriate for each of the five rating points. The evaluators then discussed various evidence that might be advanced by teachers to

indicate growth in student performance and attitude. This discussion took several directions, one of which was a treatment of the quality of evidence, for example: "Which is better evidence, improved test scores or a note from Johnny's mother saying how much better Johnny is doing in school?". Another direction was a discussion of different ways to measure attitude improvement in children, i.e. attendance and extra-curricular activities.

Much of the focus in the training was to make the evaluators comfortable with the concept of evaluation of students. Educators know when students are doing well and have positive attitudes toward learning. The purpose of the training was to provide the evaluators with examples from their own experience of this type of student evaluation, so that they could reliably assess teacher documentation of these characteristics.

Since the evaluators would also be conducting workshops to explain the system to the teachers who would be evaluated during the year, the evaluator training was seen as an opportunity to get across to the teachers through the evaluators just what constructs were being evaluated in this portion of the dialogue.

Particular emphasis was placed on the need to look at results of the evaluation to determine an accurate rating of the teacher. The previous year's results had demonstrated that high ratings were often given merely for the process of

evaluation, not necessarily the results revealed by the evaluation. A results based orientation was stressed throughout the training.

Discussion

Since measures of student performance differ from one area of instruction to another, i.e., general education, vocational education, and special education, different operational definitions of the constructs and anchors are used for each area of instruction. This definition was obtained by having experts from each area meet to develop consensus about constructs and anchors that are similar to the consensus achieved by the experts. General guidelines are necessary in order that all of the groups are in agreement on the quality which is expected at each level of the scale.

When measuring student performance and attitude on the basis of teacher documentation, it is important to remember that what is available is an indirect measure and is subject to error from various sources. This is especially important if the measure is intended to be used as a part of teacher merit evaluation - it develops its particular dimensions of sensitivity. All reasonable means must therefore be used to assure accuracy, equitability, and fairness.

Two main issues are important. The first is the modification of the scale as the result of trials and the

second involves the training of evaluators to use that scale. Within these, there is the problem of clarifying the difference between the quality of the evidence that is presented and the quality of the student performance that is inferred from the evidence. The scale was modified by using a panel of experts to discuss and define the constructs used for each item and the indicators for each scale point. In this way a consensus was achieved on how the evidence was to be evaluated. The experts then rated a given set of dialogue comments and a measure of inter-rator reliability was computed; for the dialogue as a whole, however, reliabilities weren't calculated for specific items. The experts viewed their ratings and discussed why differences in ratings occurred. When discussion was completed, the experts again rated the set of evaluator comments, and so on until consensus was achieved. The consensus was used as a benchmark for training evaluators.

As a result of this investigation, the authors believe that student performance and attitude data as presented by teachers can become a reliable part of a teacher evaluation system if the teacher evaluators are sufficiently trained in the use of the rating scale, are sensitive enough to ask the necessary "probing questions" of shy or reluctant teachers who may not have developed the showmanship of some of their peers, and if the evaluators can be trained to produce results that are reliable and valid and responsive to the <u>quality</u> of the evidence presented and not the <u>process</u> of

describing the evidence. Attention to such issues on the front end of a teacher evaluation system should pay large dividends in the useability and acceptance of the system by teachers.

References

Bartz, D., et.al., (1978). A multicomponent approach to evaluation for teacher improvement. Guide presented at the Annual Convention, National Association of Secondary School Principles (62nd, Anaheim, CA, February). Kalamazoo, MI: Kalamazoo Public Schools.

Berk, R. (1984). The use of student achievement test scores as criteria for allocation of teacher merit pay. Paper presented at the National Conference on Merit Pay for Teachers (Sarasota, FL, March 22-23, 1984).

Bolin, F. (1984). The Effective Teacher: Educational testing and measurement as a curriculum concern. Paper presented at the 68th Annual Meeting of the American Educational Research Association (New Orleans, LA, April).

Capie, W. et.al. (1980). Using pupil achievement to validate ratings of student teacher performance. Paper presented at the 64th Annual Meeting of the American Educational Research Association (Boston, MA, April).

Centra, I. (1979). Determining faculty effectiveness, assessing teaching, research, and service for personnel decisions and improvement. San Francisco, CA: Jose-Bass Publications.

Cochran, J, and Mills, C. (1983). ESL teacher effectiveness as measured by student performance on the english language proficiency test. (ERIC Document Reproduction Service No.ED246664.)

Cochran, W. (1977). Sampling techniques (3rd Edition) New York: Wiley & Soars, p. 73

Denton, J. and Norris, S. (1979). Learner cognitive achievement: A basis for establishing a student teacher's competence. (ERIC Document Reproduction Service No. 177125.)

Elliott, C., et.al. Teacher performance and pupil achievement on teacher made tests. Paper presented at the Annual Meeting of the Eastern Educational Research Association (3rd, Norfolk, VA, March).

Genck, F. (1982). Make your school system accountable. American School Board Journal. 169, (2), 34.

Good, T.L., et.al. Teaching effectiveness: Research fundings and policy implications. Paper presented for the National Institute of Education Conference on State and Local Policy Implications of Effective School Research. Washington, D.C.: Denjell Asscciates, Inc.

Hall, C. (1981). Aspects of Teacher Accountability in New Mexico. Paper presented at the Annual Meeting of the American Educational Research Association (Los Angeles, CA, April).

James, B. (1978). An examination of an approach for using pupil outcome scores to validate assessments of teacher behaviors. Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toranto, Ontario, Canada, March).

Johnson, M. (1982). Cognitive attainment of learners of student teachers across two units of study. College Station, TX: Texas A and M University

Lucio, W. (1973). Pupil achievement as an index of teacher performance. Educational Leadership. 31, (1), 71-77.

Medley, D. (1985). Issues and problems in the validation of teaching and teacher professional behaviors. Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March).

Morrison, E. (1974). Performance observation as an approach to teacher evaluation. Educational Horizons, 52, (4), 167-72.

NJEA (1979). Pupil programs as a criterion: Should it be used? NJEA Review, 52, (7), 8-10.

NJEA (1978). Pupil progress and your evaluation. NJEA Review, 52, (1), 18-21.

Ornstein, A. (1983). How good are teachers in effecting student outcomes. NASSP Bulletin, 67, (459), 73-80.

Patterson, E., and Keel (1976). Teaching effectiveness: Considerations for nursing. Journal of Continuing Education in Nursing. 7, (1), 23-32.

Regional Educational Laboratory for the Carolinas and Virginia (1970). Teachers held accountable for student learning in RELCV experiments. Automated Education Letters 5, (3), 3-5.

Reiff, J. (1981). Teacher evaluation: an essential but complicated process. Psychology: A Quarterly Journal of Human Behavior. 18, (1), 24-27.

Scheim, M. (1985). Student achievement as a measure of teacher effectiveness. Journal of College Science Teaching. 14, (6), 471-74.

Soar, R., and Soar, R.M. (1975). Problems in using pupil outcomes for teacher evaluation. Washington, D.C.: National Education Association.

Footnotes

[1]The scale was hypothesized to contain a 1-5 conterminous points where 1 = no evidence or description, 2 = vague or unclear statements, 3 = clear statements with at least 1 example, 4 = clear statements with several examples, and 5 = clear statements with multiple examples.